DOCUMENT RESUME

ED 232 041 TM 830 455

AUTHOR Holley, Freda
TITLE Collaboration between Testing and Content Specialists to Improve Teacher-Made Tests. Publication #82.53.
INSTITUTION Austin Independent School District, Tex.
PUB DATE Apr 83
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; *Inservice Teacher Education; Item Banks; School Districts; Skill Development; *Teacher Improvement; *Teacher Made Tests; *Test Construction; *Testing; Test Use
IDENTIFIERS *Austin Independent School District TX

ABSTRACT

The Austin Independent School District (AISD) became interested in evaluation and testing skills of teachers when district and school averages on competency ratings in the area were among the lowest since 1979. School districts, as well as teacher preparation programs, should devote serious attention to the improvement of these teacher skills. In the AISD, three initial activities were undertaken to improve teacher competencies in student evaluation: (1) evaluating teacher tests to generate development needs and identify resources for staff development; (2) creating test item bank projects for staff development in a practical format; and (3) long range staff development by gathering testing information and conducting training in test construction, test interpretation, test use, and student test anxiety. Staff development in testing and evaluation should be a districtwide priority. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233041

COLLABORATION BETWEEN TESTING AND CONTENT SPECIALISTS TO IMPROVE TEACHER-MADE TESTS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Freda M. Holley
Austin Independent School District

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
F. Holley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of
the American Educational Research Association,
Montreal, Canada, April, 1983

Publication #82.53

TM 830 433

COLLABORATION BETWEEN TESTING AND CONTENT SPECIALISTS TO IMPROVE TEACHER-MADE TESTS

Freda M. Holley

Austin Independent School District

We in the Austin Independent School District became interested in evaluation and testing skills of teachers when we began compiling district and school averages on competency ratings given to teachers on the annual teacher evaluation form. When ratings on 46 competencies were examined, these were among the lowest rated in 1979:

- Prepares appropriate test and evaluation activities to measure student learning.
- Interprets own tests and evaluation activities accurately.
- Documents student progress effectively.

In 1980, 1981, and 1982, this continued to be the case (Totusek, 1982).

These competencies had initially been included in Austin Independent School District's new professional evaluation system because teachers, parents, supervisors, and administrators had given these competencies the highest ratings of importance. Therefore, these low ratings seemed worthy of real concern.

Later national research findings came to our attention which contributed to our estimate of the importance of teacher testing competencies. For example, Ron Edmunds' description of effective schools noted, "Pupil progress is monitored frequently. The teachers use frequent classroom tests to monitor student progress toward instructional objectives" (Edmunds, 1979). A study conducted by the

Center for the Study of Evaluation (1982) found that large amounts of time, as much as 10% at the high school level, went to testing; fully three-quarters of this testing time was consumed with teacher-developed tests. Yet, the study pointed out that while teachers relied mostly on their own tests for decision-making, administrators rarely reviewed or evaluated these tests and tended to focus their attention on state or district-mandated tests (Center for the Study of Evaluation, 1982).

All of these findings suggest that school districts as well as teacher preparation programs should devote serious attention to the improvement of teacher skills in this area. In Austin ISD, some initial steps in this direction have been taken. This paper outlines three activities undertaken.

Evaluating Teacher Tests

First, administrators had to look at exams. To initiate this, the director of secondary education required junior high principals to collect and review samples of their teachers' tests. Principals found this activity revealing. Teachers became aware of the importance of their exam quality and appearance.

The following year both junior and senior high principals collected three exams from each teacher. A study of these tests was contracted to local university staff (Measurement and Evaluation Center, University of Texas at Austin). They scored these exams using the checklist and criteria shown as attachments 1 to 3. This study produced a districtwide summary addressed to principals. The summary showed that:

- There was good variety in the types of items on the tests. (See Table 1, Attachment 4.) There was not too much reliance on one item type.

- Overall, the tests were good. (See Table 2, Attachment 5.) However, there was also room for improvement. For example, 30% of the directions given were not fully adequate.
- Also, some items at every level of Bloom's Taxonomy appeared on the tests. (See Table 3, Attachment 6.) But the percentage of items at the higher level was low.

The study concluded that staff development would be beneficial.

A script for a faculty meeting was prepared to assist principals in returning the exams to teachers. Each teacher received one exam with a feedback sheet. The script also helped principals to explain the purposes of the study and to lead the faculty in generating needs and identifying resources for staff development on testing competencies. (See Attachment 7.)

A number of districtwide staff development sessions on testing have occurred. However, an improvement need of the kind indicated by our current data can only be remedied by a long range intensive effort.

Test Item Bank Projects

A more immediate instrumental way to improve teacher tests is to work with them to create item banks. AISD's social studies and science instructional coordinators, with some support from our Office of Research and Evaluation, have been working with their teachers for three years to develop item banks. In social studies there are currently banks for American History and World History. The World History bank is organized by historical periods such as "Prehistory." Each item is currently recorded on a card with base information on item form, behavior level (from Bloom Taxonomy), scholastic level, state curriculum goals and objectives and item source. Eventually both item banks can be computerized for efficient word processing.

Item banks are actually useful in three ways. First, the teachers participating in the development receive a great deal of staff development in a practical, acceptable format. Through direct and immediate application teacher competency is likely to increase. Second, student evaluation is likely to be higher in quality through better tests. And, third, teacher time will be saved from test development to test production.

Long-range Staff Development

The social studies coordinator initiated a staff development effort to improve social studies teachers' testing competencies following the receipt of the first year's data showing inadequacies in that area. This involved the development of a notebook with testing information such as test company brochures on definitions and test construction, training sessions with ORE staff on test construction and how to deal with student test anxiety, and sessions led by the coordinator on interpreting and using standardized test data.

Conclusion

Each of the efforts described above is seen as a possible way to impact teacher competencies for student evaluation. Undoubtedly, they have been beneficial. The fact remains, however, that the teacher evaluation data continues to show that these competencies are low. My experiences indicate to me that we are unlikely to make a significant difference in this area unless it becomes a real districtwide priority. Unfortunately, district priorities tend to focus on desegregation, reorganization, budgets, bond issues, and other non-instructional matters. Nevertheless, the promise of payoff, however limited, leaves this an area to which we still try to devote whatever attention we can.

EXAM RATING SHEET

Attachment 1

Teacher ID No. _____

Subject _____

You will find below a summary of the characteristics of the attached test based on a sampling of the content of the test. This summary includes a consideration of appearance, directions, specific types of items, freedom from item bias, and levels of Bloom's taxonomy. The criteria used in describing the test are derived from generally accepted standards of test construction.

Feedback on Teacher-Made Tests

	Good	Fair	Marginal
Appearance (legibility, organization, neatness)	[ ]	[ ]	[ ]
Directions:	**Always**	**Sometimes**	**Seldom**
Present for each section	[ ]	[ ]	[ ]
	Good	**Fair**	**Marginal**
Clarity	[ ]	[ ]	[ ]

Analysis of Items

The desirable characteristics for each item type are listed below. If the characteristic could not be rated, it was left blank.

Multiple Choice

	Usually	Sometimes	Rarely
The phrase that introduces the item (stem) clearly states the problem.	[ ]	[ ]	[ ]
A single idea is tested in each item.	[ ]	[ ]	[ ]
Avoids use of negatives and/or double negatives in item.	[ ]	[ ]	[ ]
The length of response options in each item is consistent.	[ ]	[ ]	[ ]
Grammatical cues to the correct answer are absent.	[ ]	[ ]	[ ]
The response options are indented and in column form.	[ ]	[ ]	[ ]
Avoids use of "All of the above," "None of the above."	[ ]	[ ]	[ ]

True-False

Avoids use of "always" and "never" in the items.	[ ]	[ ]	[ ]
Avoids use of negatives in stating the items.	[ ]	[ ]	[ ]
A single idea is tested in each item.	[ ]	[ ]	[ ]
Items are brief and to the point.	[ ]	[ ]	[ ]

Completion

Only one blank space is included in each item.	[ ]	[ ]	[ ]
Blank spaces occur at the end of the sentence.	[ ]	[ ]	[ ]
Grammatical cues to the correct answer are absent.	[ ]	[ ]	[ ]
Items are worded to limit the range of possible answers.	[ ]	[ ]	[ ]

Exam Rating Sheet Continued

Matching

	Usually	Sometimes	Rarely
Directions state whether options can be used more than once.	[ ]	[ ]	[ ]
More options were present than items.	[ ]	[ ]	[ ]
Longer phrases are used as items, and shorter phrases are used as options.	[ ]	[ ]	[ ]
All items are of the same class and all options are of the same class (e.g., a list of events to be matched with a list of dates).	[ ]	[ ]	[ ]

Essay

The length of the answer desired is specified for each item.	[ ]	[ ]	[ ]
The relative emphasis given to each question is indicated (e.g., time to be spent or points assigned).	[ ]	[ ]	[ ]
Items are stated precisely, clearly focusing on the desired answer.	[ ]	[ ]	[ ]
All students are required to answer the same questions.	[ ]	[ ]	[ ]

Item bias

Items are stated in a manner that avoids bias in regard to sex.	[ ]	[ ]	[ ]
Items are stated in a manner that avoids bias in regard to race.	[ ]	[ ]	[ ]

Coverage of Bloom's taxonomy of the Cognitive domain

	Many of the Items	Some of the Items	Very Few of the Items
Items cover the knowledge level (e.g., rote memorization; basic facts)	[ ]	[ ]	[ ]
Items cover the comprehension level (e.g., understanding; ability to state in own words).	[ ]	[ ]	[ ]
Application (e.g., apply knowledge to a new situation).	[ ]	[ ]	[ ]
Analysis (e.g., ability to break down into component parts).	[ ]	[ ]	[ ]
Synthesis (e.g., ability to bring ideas together to create something original).	[ ]	[ ]	[ ]
Evaluation (e.g., ability to make decisions).	[ ]	[ ]	[ ]

GLOSSARY OF RATING STANDARDS

1. Appearance (legibility, organization, neatness):

rate the overall quality of the appearance of the test, taking into account how easy it is to read the test, its legibility, neatness, the organization or layout of the sections of the test, and duplication quality.

Good = test meets the appearance criteria listed above
Fair = test meets several, but not most, of the criteria above
Marginal = test meets few or none of the criteria above

2. Directions:

A. Present for each section — scan through the entire test to identify each separate group of item types; check to see if directions are written out

Always = every separate group of items is accompanied by instructions
Sometimes = directions are usually present for each section
Seldom = written directions are present for only one or two sections, or are not present at all

B. Clarity — carefully read the directions that are present for each section of items on the test; rate the directions in terms of how clearly they instruct the student as to what he/she is to do in answering the questions or responding to items.

Good = directions are clear and unambiguous; student knows what to do
Fair = there is some uncertainty as to how to answer the items
Marginal = it is difficult to know what to do to answer the items, or there are conflicting instructions; vague or ambiguous

3. Analysis of Items:

Usually = the items that were sampled met the standard or guideline the majority of the time
Sometimes = the standard or guideline was met less than half of the time for the items sampled
Rarely = only one or two of the items sampled (or none) met the standard

4. Coverage of Bloom's Taxonomy:

Many of the Items = the majority of the items sampled measured this level
Some of the Items = less than half of the items sampled measured this level
Very Few of the Items = only one or two or none of the items sampled measured this level

NOTE: If you are unable to rate any particular characteristic of the test, leave the response categories blank.

TABLE 1

ITEM TYPES INCLUDED ON THE TESTS

ITEM TYPE	YES	NO
Multiple Choice	49%	51%
True/False	31%	69%
Completion	45%	55%
Matching	42%	58%
Essay	41%	59%

n = 1,978 tests

TABLE 2

GLOBAL EVALUATIONS OF THE TEACHER-MADE TESTS

CRITERION	GOOD	FAIR	MARGINAL
Quality of Test Appearance	69%	25%	6%
Adequacy of test Directions	70%	19%	11%
Performance in Meeting Item Writing Standards	59%	38%	3%

n = 1,978 tests

TABLE 3

COVERAGE OF THE LEVELS OF BLOOM'S TAXONOMY

LEVEL	MANY ITEMS	SOME ITEMS	VERY FEW ITEMS
KNOWLEDGE	77%	15%	8%
COMPREHENSION	20%	37%	43%
APPLICATION	13%	33%	54%
ANALYSIS	4%	23%	73%
SYNTHESIS	1%	10%	89%
EVALUATION	1%	14%	85%

n = 1,978 tests

Attachment 7

DEPARTMENTAL STUDENT EVALUATION QUESTIONNAIRE

_____ Department

_____ High School

_____ Date

A. Areas of strength in student evaluation and testing in this department:

1.

2.

3.

4.

5.

B. Areas that we can improve in student evaluation and testing in this department:

1.

2.

3.

References

Center for the Study of Evaluation. CSE analyzes school testing programs. CAPTRENDS, Clearing House for Applied Performance Testing, Northwest Regional Educational Laboratory, May 1982, Vol. 7(3), 2-3.

Edmunds, R. Effective schools for the urban poor. Educational Leadership, October 1979, pp. 2-3.

Measurement and Evaluation Center. Project summary report: Analysis of AISD Secondary School Teacher-Made Tests. Austin, Texas: Measurement and Evaluation Center, University of Texas at Austin, May, 1982.

Totusek, P. 1981-82 Final Report: Personnel Evaluation Systems. Austin, Texas: Austin Independent School District, Office of Research and Evaluation, publication number 81.30, 1982.